Supplementary information: Fast and sensitive mapping of bisulfite-treated sequencing data.

Christian Otto [1,2], Peter F. Stadler [1–6], Steve Hoffmann [1,2*]

July 10, 2012

[1]Bioinformatics Group, Dept. of Computer Science, University of Leipzig, Germany
[2]LIFE - Leipzig Research Center for Civilization Diseases, Universität Leipzig, Germany
[3]RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology, Leipzig, Germany
[4]Santa Fe Institute, Santa Fe, New Mexico, USA
[5]Department of Theoretical Chemistry, University of Vienna, Austria
[6]Max-Planck-Institute for Mathematics in Sciences, Leipzig, Germany

List of Figures

List of Tables

*to whom correspondence should be addressed

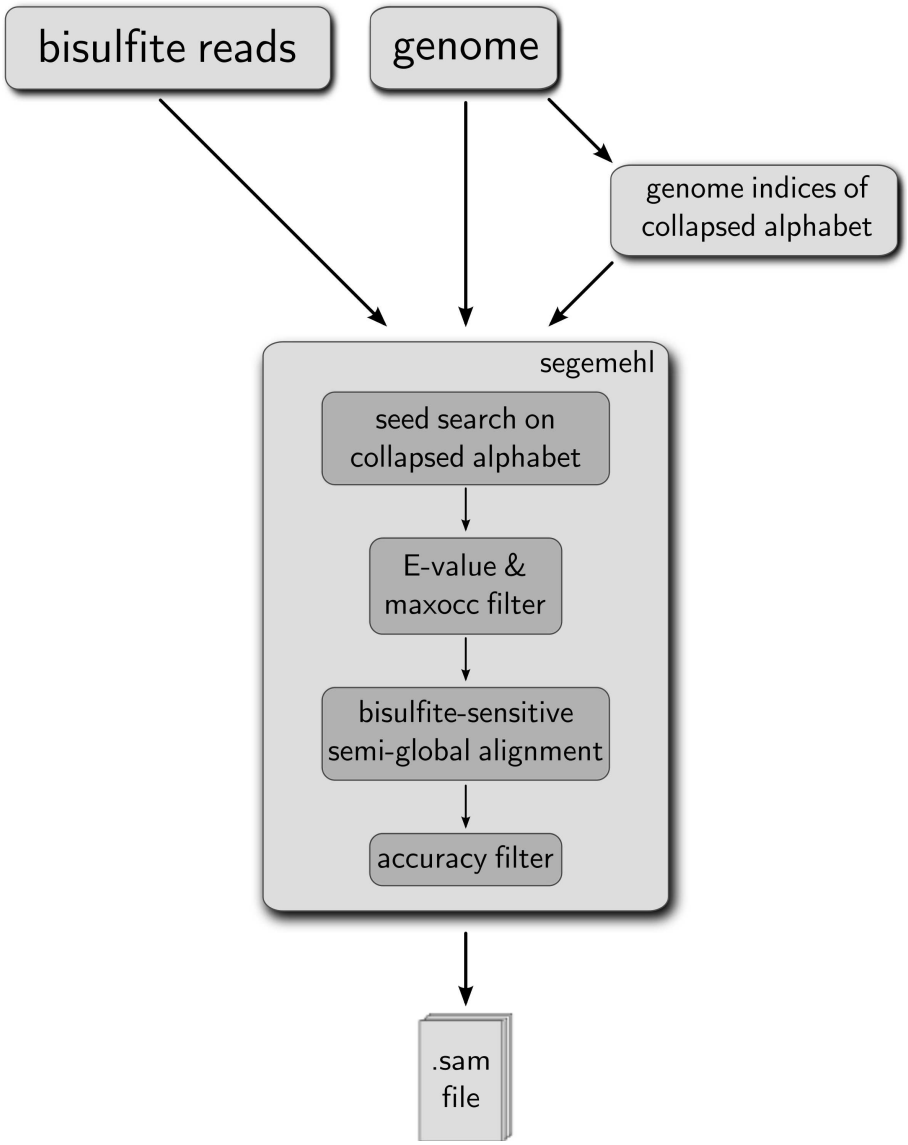
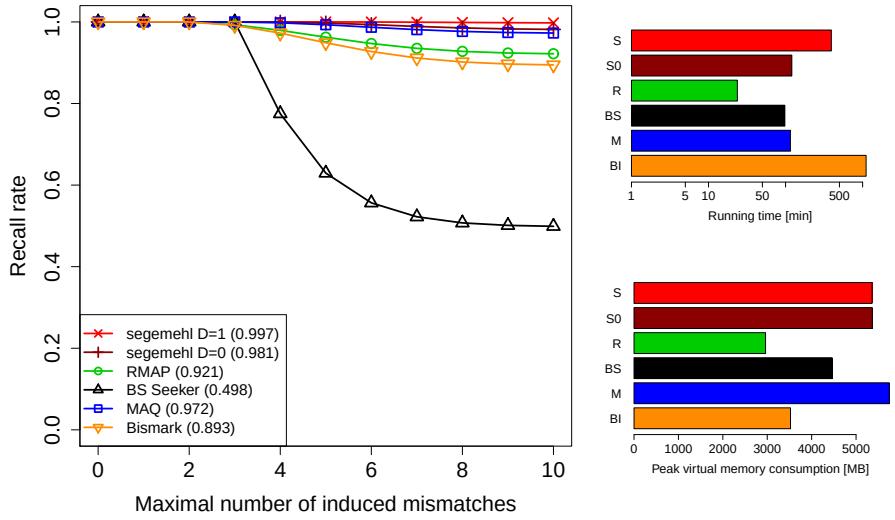
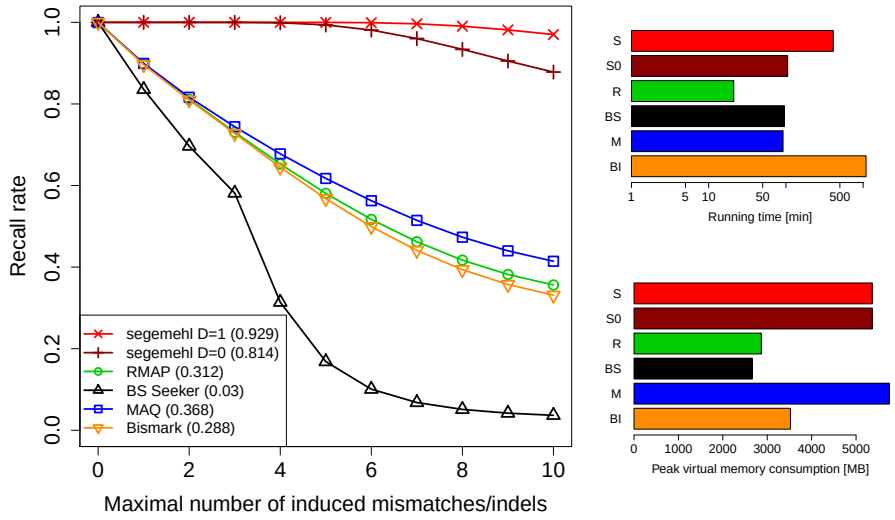


Figure 1: **Bisulfite mapping with `segemehl`**. As input, `segemehl` requires reads in fastA or (gzip'd) fastQ format, the genome sequence, and two enhanced suffix arrays (ESA) created for the genome sequence with collapsed alphabet, i.e., C/T and G/A conversion. Then, `segemehl` performs the seed search on the collapsed alphabet by use of the ESAs with greedy matching statistics. It further discards seeds with high expectation value and those that occur very often in the genome. Both filtering steps are controlled by user-defined threshold parameters. In the following, each reads is extended to a semi-global alignment using Myers bit-vector algorithm [1] which was further extended to support the IUPAC nucleotide code. Hence, asymmetric bisulfite-related mismatches, i.e., mismatches introduced by the treatment with sodium bisulfite, are treated as match and not penalized. In order to prevent spurious hits, read alignments with insufficient accuracy (fraction of matches) are rejected. In the end, `segemehl` produces an output file in standardized SAM format for which many post-processing tools are available such as samtools [2] and Picard (`http://picard.sourceforge.net/`)



(a)



(b)

Figure 2: **Performance evaluation on artificial datasets with 5% mismatches or 10% mismatches + indels.** The benchmarks assessed the performance of `segemehl` with D=1 (in red) and D=0 (in dark red), `RMAP` (in green), `BS Seeker` (in black), `MAQ` (in blue), and `Bismark` (in orange) in terms of recall rate, running time (in user mode), and peak virtual memory consumption by mapping 10 million artificial bisulfite reads to a 200 MB large random reference. Furthermore, (a) mismatches at a rate of 5% or (b) mismatches + indels at a rate of 10% were randomly introduced into the bisulfite reads. The recall rate is the relative number of mapped reads where the score of the best alignment is found to be unique and the original position on the artificial reference was recovered correctly. The recall rate was estimated on subsets of the artificial reads with limited number of introduced mismatches or mismatches+indels. The overall recall rate of each program with the entire query dataset is given in the legend. Note that the preprocessing time is not included in the measurement.

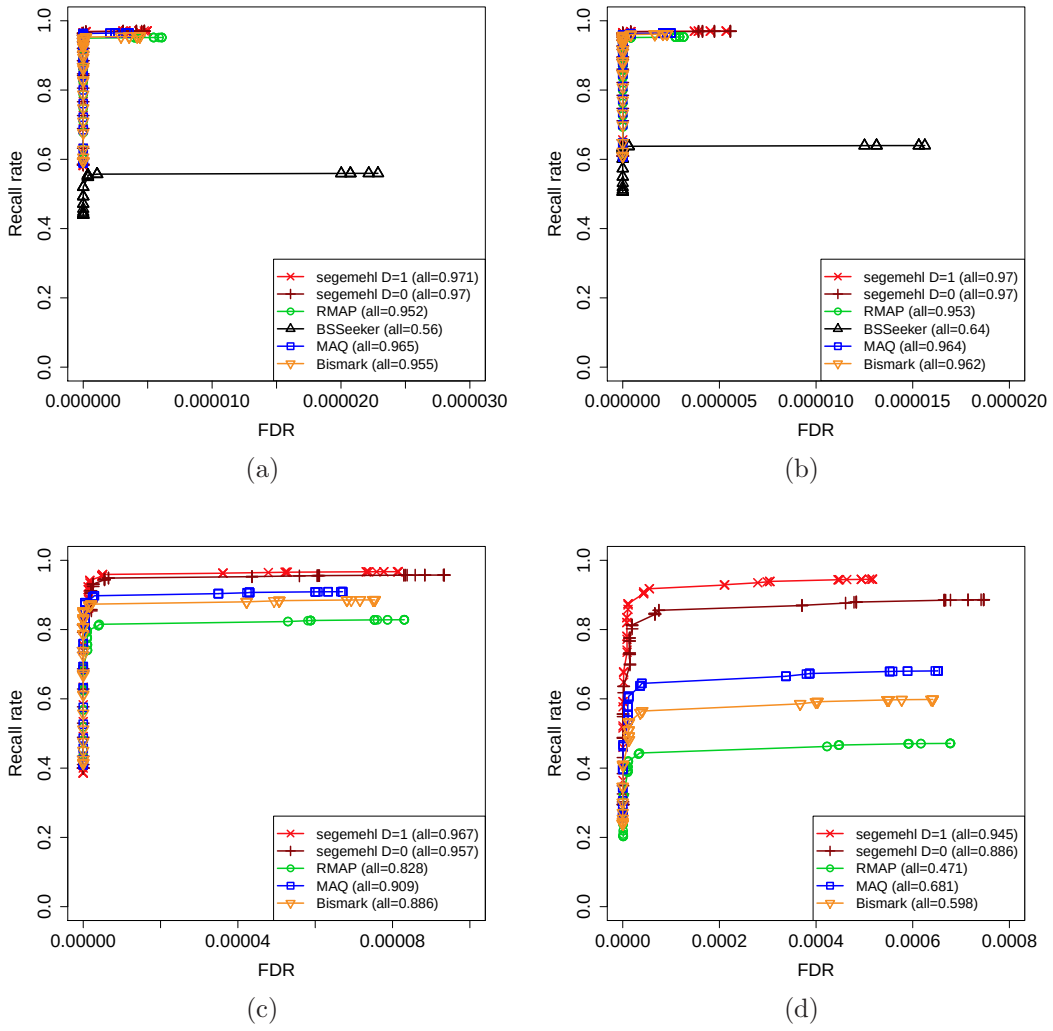


Figure 3: **Performance in methylation calling benchmarks with 5% mismatches or 5%, 10%, or 15% mismatches + indels.** Recall rate as function of FDR after evaluating the performance in methylation calling using the mapping output of `segemehl` with D=1 (in red) and D=0 (in dark red), `RMAP` (in green), `BS Seeker` (in black), `MAQ` (in blue), and `Bismark` (in orange). We therefore mapped 2.5 million artificial bisulfite reads, containing (a) mismatches or (b-d) mismatches + indels at a rate of 5% (b), 10% (c), or 15% (d), with each program to a random 10 MB large reference sequence. Ambiguously mapped reads were discarded. For each cytosine on the reference covered by at least 5 reads on the same strand, the methylation state was called using a simple majority voting approach. Recall rates and FDRs were then estimated at different score cutoffs, see Methods section for further details. The peak recall rate with each program is given in the legend.

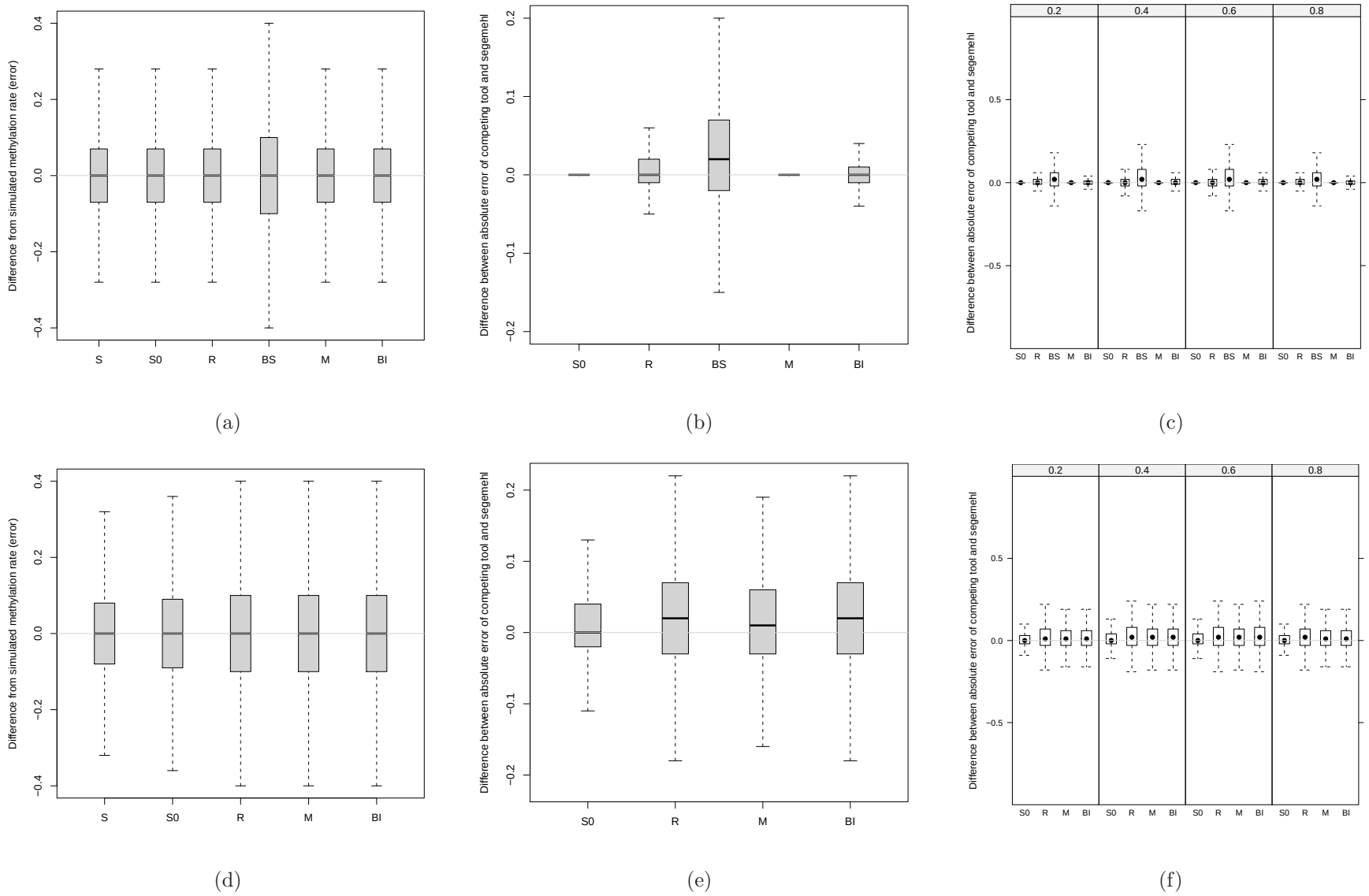


Figure 4: **Performance in methylation rate benchmarks of sufficiently (≥10-fold) covered cytosines with 5% or 15% mismatches.** Performance in recalling the simulated methylation rates using the mapping output of `segemehl` with D=1 (S) and D=0 (S0), `RMAP` (R), `BS Seeker` (BS), `MAQ` (M), and `Bismark` (BI). Five million artificial bisulfite reads, containing 5% (a-c) or 15% (d-f) mismatches, were generated from a random 10 MB reference sequence with simulated methylation rates (20%, 40%, 60%, or 80%). The expected coverage of the reference was 20. Subsequently, reads were mapped back to the reference with each tool. Ambiguously mapped reads were discarded. The methylation rate of each cytosine covered by at least 10 reads was estimated as fraction of non-converted over the sum of non-converted and converted bases at this position. The differences from the simulated methylation rates (errors) are illustrated as box-whisker plots (a,d). To compare the other tools directly with segemehl (D=1), at any given cytosine that was covered at least 10-fold **by both** aligners, the absolute error of segemehl's estimate was substracted from the absolute error of the competing tool (b,e). Hence, values above zero denote positions where the estimated methylation rate with the mapping output of segemehl is more accurate. The analysis was repeated for the four simulated methylation rates separately (c,f).

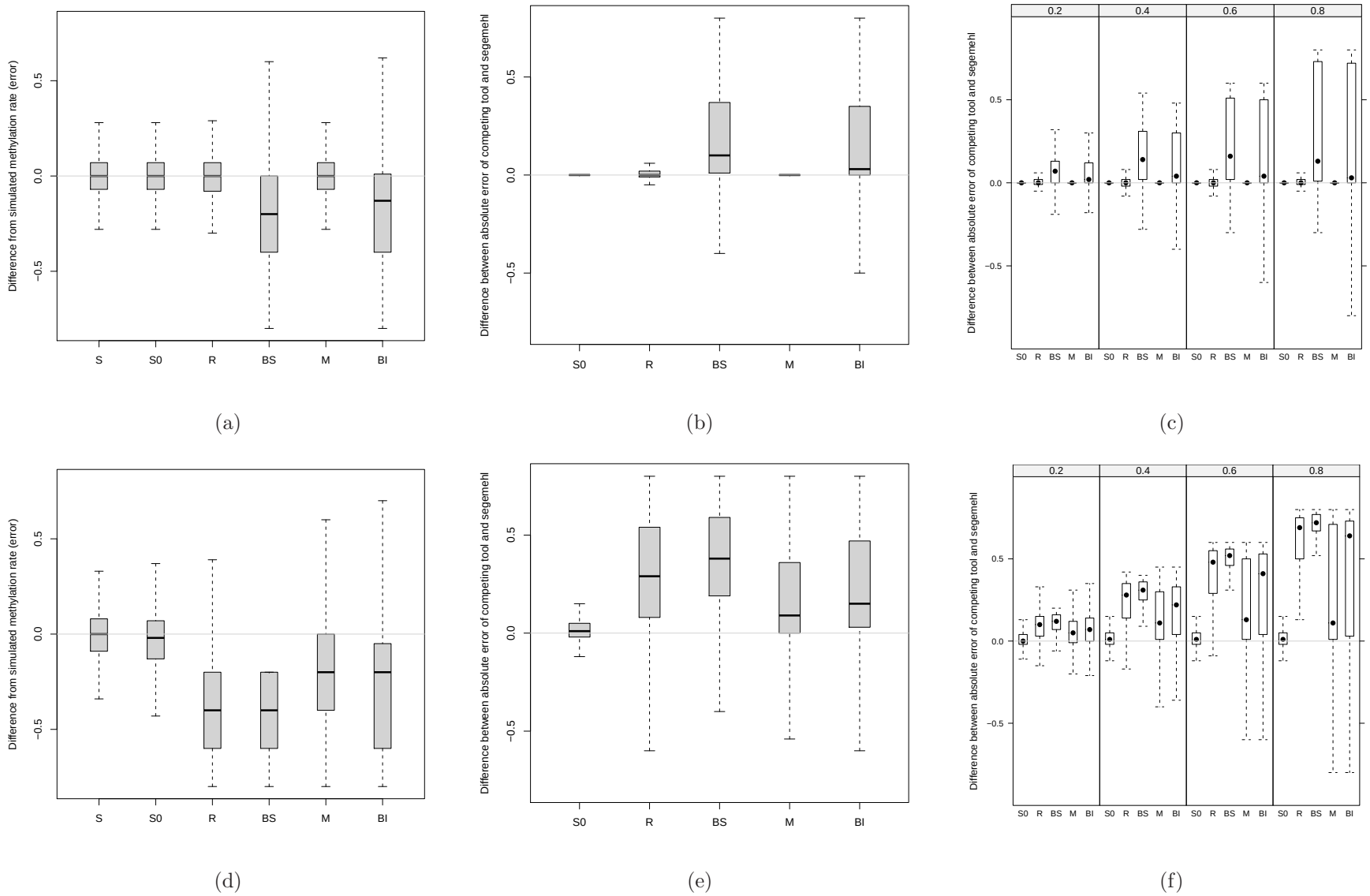


Figure 5: **Performance in methylation rate benchmarks of all cytosines with 5% or 15% mismatches.** Performance in recalling the simulated methylation rates using the mapping output of `segemehl` with D=1 (S) and D=0 (S0), `RMAP` (R), `BS Seeker` (BS), `MAQ` (M), and `Bismark` (BI). Five million artificial bisulfite reads, containing 5% (a-c) or 15% (d-f) mismatches, were generated from a random 10 MB reference sequence with simulated methylation rates (20%, 40%, 60%, or 80%). The expected coverage of the reference was 20. Subsequently, reads were mapped back to the reference with each tool. Ambiguously mapped reads were discarded. The methylation rate of each cytosine covered by at least 10 reads was estimated as fraction of non-converted over the sum of non-converted and converted bases at this position. The estimated methylation rate at insufficently covered sites was set to zero. The differences from the simulated methylation rates (errors) are illustrated as box-whisker plots (a,d). To compare the other tools directly with segemehl (D=1), at any given cytosine that was covered at least 10-fold **by one** of the two aligners, the absolute error of segemehl's estimate was substracted from the absolute error of the competing tool (b,e). Hence, values above zero denote positions where the estimated methylation rate with the mapping output of segemehl is more accurate. The analysis was repeated for the four simulated methylation rates separately (c,f).

Table 1: **Performance evaluation on good-quality real-life dataset.** The tests assessed the performance of `segemehl` with D=0 or D=1 (`-F 1`, `-H 1`, `-A 70`), `BS Seeker` (`-t N`, `-e 80`, `-m 3`), `RMAP` (`-B`, `-m 20`), `MAQ` (`-M c`, `-n 3`, `-e 500`), and `Bismark` (`--directional`, `-n 3`, `-e 500`) by mapping a part of the recent whole genome shotgun bisulfite sequencing dataset of the human induced pluripotent stem cell line derived from foreskin fibroblasts FF-iPSC 19.11 (published by [3]) against the Human genome in terms of running time (in user mode), peak virtual memory consumption, and fraction of unique best mapped reads (overall or subdivided by the maximal number of mismatches + indels in the alignment). Note that last measure only considers read mappings where the score of the best alignment is found to be unique. The best value in each measure, e.g., lowest running time, lowest memory consumption, or highest number of unique best mapped reads, is printed in boldface. As input dataset, we used one tenth of the real-life dataset SRR094462, i.e., 17'480'543 reads of length 85 nt. This dataset, particularly the base sequence qualities and sequence quality scores, is considered as a good-quality illumina dataset according to the FastQC toolkit. Note that the preprocessing time is not included in the time measurement. Details on the selected parameters of each program are given in the Methods section.

	running user time (min)	memory[1] (MB)	mismatches+indels = 0	≤ 1	≤ 2	≤ 3	≤ 4	≤ 5	≤ 10	max
`segemehl (D=1)`	765	75001	64.4	**76.9**	**80.7**	**82.6**	83.9	**84.8**	87.6	89.4
`segemehl (D=0)`	257	75001	64.4	76.9	80.7	82.6	**83.9**	84.8	**87.6**	**89.6**
`BS Seeker`	**239**	12824	64.0	76.1	79.6	81.5	81.5	81.5	81.5	81.5
`RMAP`[2]	1296	**8123**	**64.4**	76.5	80.0	81.8	82.9	83.7	86.2	88.1
`MAQ`[23]	24822	749	63.0	74.6	77.8	79.3	80.3	81.0	83.2	85.0
`Bismark`	2197	14649	63.9	75.7	79.1	80.7	81.8	82.6	85.0	87.1

[1]Virtual memory consumption shown while the required physical memory considerably less. For example, `segemehl` uses only around 52 GB of physical memory. [2]`RMAP` and `MAQ` do not provide multi-threading. [3]The results of `MAQ` are estimated using 5% of the data.

References

[1] G Myers. A fast bit-vector algorithm for approximate string matching based on dynamic programming. *Journal of the ACM (JACM)*, 46(3):395–415, 1999.

[2] H Li, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–9, Aug 2009.

[3] R Lister, M Pelizzola, Y S Kida, R D Hawkins, J R Nery, G Hon, J Antosiewicz-Bourget, R O'Malley, R Castanon, S Klugman, M Downes, R Yu, R Stewart, B Ren, J A Thomson, R M Evans, and J R Ecker. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, 471(7336):68–73, Mar 2011.